U^2RPSA: Leveraging Unsupervised Learning for Water Main Decision Support

Neh Majmudar[1] , **Justin Gbadamassi**[2] , **Anita Raja**[1,2]
[1]City University of New York, Graduate Center
[2]City University of New York, Hunter College
neh.majmudar@gradcenter.cuny.edu, abdoud.gbadamassi20@myhunter.cuny.edu,
anita.raja@hunter.cuny.edu

Abstract

The urban water supply system is crucial for city life, yet it remains vulnerable to a range of disruptions, particularly in densely populated areas. The motivation for this work is to design decision-support algorithms for early prediction of water main breaks and to prevent the potential damage to life and property. We present a comprehensive approach that uses statistical learning techniques and Bayesian networks involving three key steps: unsupervised learning of a Bayesian network structure to handle uncertainty in data and action outcomes, water main break risk prediction using machine learning, and survival analysis to estimate the probability of a pipe's longevity. Utilizing a publicly available dataset, we provide an initial evaluation of our approach showing that it outperforms a state-of-the-art model while providing a holistic understanding of pipe breakage dynamics for infrastructure maintenance.

1 Introduction

The urban water system is vital for city life but susceptible to disruptions, like the water main breakage above a Manhattan subway station in August 2023 that affected the commute of 300,000 passengers and led to costly cleanup efforts [Nolan, 2023]. A month later, Storm Ophelia caused additional flooding, further disrupting lives and damaging property [Ley, 2023]. NYC transit authorities and the Metropolitan Transportation Authority (MTA) have acknowledged the challenges posed by an aging infrastructure, which includes thousands of pipes, some of which date back to the 1890s, underscoring the pressing need for regular maintenance and replacement.

Designing decision-support algorithms for modeling, predicting, and preventing water main breaks and their cascading consequences across the infrastructure network is the motivation for this work. Specifically our twin goals are: (a) early risk prediction of water main breakage which in turn allows for (b) mitigating actions that prevent the breaks and associated damage. We propose the building blocks for an AI-powered framework for recommending areas needing attention, such as locations with missing data and potential water main break locations in city infrastructure. Our proposed model will use historical data to predict and manage these incidents effectively, offering cost-saving strategies for network maintenance. Critical features for water main break prediction identified in the literature include pipe condition, age, size, material, and environmental factors like traffic, location, soil, groundwater, and climate [Kumar *et al.*, 2018; Kabir *et al.*, 2016]. The work described in this paper aims to build a sustainable system that not only relies on the prediction of the risk of a pipe breaking but also investigates ways to identify and where possible, mitigate the potential causes for the breakage. In this study, we present U^2RPSA, a pipeline that starts by understanding the dependencies between the factors involved in water main pipe breakage using Unsupervised Bayesian Learning under Uncertainty (U^2). In the next step, the probability of a pipe breaking is determined based on identifying patterns within the data using Risk Prediction (RP). In the last step, Survival Analysis (SA) aims to estimate the probability of a pipe surviving after a certain time in the future. All three steps together enable the decision support system to prioritize pipe maintenance in order of urgency (highest risk of breakage) and to take the mitigation steps early enough to prevent breakage.

2 Related Work

Risk Prediction: The prediction of water pipe breakage relies on critical features such as the history of breaks, pipe conditions (age, size, material, and water pressure [Kumar *et al.*, 2018; Demissie *et al.*, 2017; Á. Martínez-Codina and Garrote, 2016], and environmental factors [Kabir *et al.*, 2016] like traffic patterns and soil conditions. Various classical machine learning (ML) techniques, have been explored for this purpose. Bayesian networks [Tang *et al.*, 2019], known for modeling complex dependencies and uncertainties, employ directed acyclic graphical (DAG) models, integrating multiple data sources and expert judgments, offering interpretable graphical structures, and accommodating interventions [Koller and Friedman, 2009]. Prior research has employed dynamic Bayesian networks [Demissie *et al.*, 2017] to assess the impact of time-dependent factors on pipe failures, emphasizing the significance of time in prediction models. Feature selection, as demonstrated by Omar et al. [Omar *et al.*, 2023] plays a crucial role in enhancing predictive accuracy, focusing on factors like pipe age, material, condition score, and criticality. **Survival Analysis** There is also prior work in predicting pipe breaks using Survival Analysis. Survivability refers to the estimation of the probability of an event occurring after a certain time has elapsed. Somek [Kimutai *et al.*, 2015] compare various statistical regression models to estimate the survival function. Kabir et al. [Kabir *et al.*, 2016] uses the Bayesian Model Averaging framework on Survival Analysis models, by taking into account the time-dependent covariates. They develop the survival curves with sequential parameters, that are observed by dividing the years into discrete periods.

However, these related works only focus on specific tasks (i.e. Risk Prediction or Survival Analysis) with respect to pipe breaks. Also, they do not delve into the reasons for the pipe breaks. Our approach aims to address this knowledge gap by predicting the status and estimating the survival chances of risk-prone pipes. We also investigate the causes for pipe breaks as part of our overall goal to pursue mitigation strategies, thus providing a framework for the holistic understanding for prediction and prevention of water main breaks.

3 Approach

Our proposed framework for prediction and prevention of pipe breaks involves 3 steps. Step 1 is unsupervised learning of a Bayesian network (BN) structure; Step 2 involves risk prediction using Machine Learning, and Step 3 deals with Survival Analysis that is leveraged for estimation of the survival probability of a pipe. We will now elaborate upon our methodology in detail.

Step 1: Learning a BN structure captures the statistical correlations and relationships between features in the data representing the city's water infrastructure with minimal training input. This automated approach used minimal expert guidance in the form of a temporal blacklist. Unlike prior works [Kumar *et al.*, 2018], where a set of ML techniques requiring all features for final label prediction are applied and compared to determine the best approach, our approach can achieve water main predictions just using Markov blanket in the testing phase. The methodology implicitly reveals the most significant features in the dataset through its Markov blanket representation. We employed structure learning algorithms to determine the edges in the Bayesian network, focusing on score-based methods like the Bayesian Information Criterion in Equation 1 in our study.

$$BIC = logP(D|G) + d/2log(N) \tag{1}$$

Here, D is the data, G is the structure, d is the number of free parameters in the network, and N is the size of the dataset on which the learning is being conducted. The learned model approximates a graphical model, serving as a map of the total distribution of the process. The pipeline consists of six steps, including input processing, data transformation, visualization, structure learning, structure averaging, and prediction and analysis [Mallia, 2023].

Step 2: Risk Prediction To visualize the results of the above Step 1, we utilize a tool that highlights the probabilistic relationship modeling offered by Bayesian Networks. While several programs exist for BN modeling [Uusitalo, 2007] and causal discovery [Glymour *et al.*, 2019], we use "BN Inspector", a bespoke tool [Mallia, 2023] designed by our research team, that has the following basic functions: (1)generates a dynamic plot of the BN currently being utilized, which updates to reflect the Markov Blanket of the currently selected target outcome, as well as what variables are being used in evidence for examining conditional distributions. (2) facilitates analysis of marginal versus conditional probabilities for a variable: the user enters available evidence in the form of a logical expression and observes the impact on the distribution for a variable as discovered via an inference algorithm [Scutari, 2010]. Once the relationships between features are visualized and understood using our BN inspector tool, the next stage is to utilize the features for modeling the prediction of a pipe break. We perform data pre-processing and merging on datasets that could represent a range of databases including water main pipe distribution, historical break data, data on factors that could affect water main breaks such as traffic, soil type, pressure zones, tree root growth etc. Section 4 describes the datasets we used for evaluation in more detail. Then, we incorporate Analysis of Variance significance tests (ANOVA) to calculate the significance of features for the target outcome viz., $Pipe\_Break$. We do extensive cross-validation using Optuna [Akiba *et al.*, 2019] for hyperparameter selection. Ten-fold cross-validation involved multiple trials of six classification models, including Logistic Regression, Naive Bayes, Decision Tree, Random Forest, K-Nearest Neighbors, and XGBoost. The evaluation results of these models are compared with the state-of-the-art results.

$$S(t) = P(T > t) \tag{2}$$

$$S(t) = \int_t^\infty f(u)\,du = 1 - F(t) \tag{3}$$

Step 3: Survival Analysis In addition to identifying which pipes will break, we are also interested in determining when they will break with a high level of confidence. We use Survival Analysis (SA) to do this analysis. SA uses a set of statistical methods to analyze time-to-event data. Proportional hazard models (PHM) are a class of statistical models that were used in this study. They determine the relationship between independent variables and the hazard function over time. Cox models [Cox, 1972], rely on the assumption that the relative hazard of an event between two groups remains constant over time. Using features like $'Pipe\_Installation\_Year'$, we calculated the duration until breakage or censoring. The survival probability can be estimated by using the Survival Function, as given in Equation 2, where T is the time to death, and $S(t)$ is the chance of a pipe surviving after time t. As seen in Equation 3, the function can also be modeled using the hazard function, which is the probability of occurrence of the event at T=t, assuming that the event has not occurred up through time t. Here, $f(u)$ is the hazard function, $F(t)$ is the cumulative hazard function from 0 to time t and censored data means that either the pipe did not break, or the study didn't show that the pipe broke. Model selection was performed for Cox PHM, with the penalization factor as the parameter. We also use XGBoost's(XGB) model with Survival Embeddings

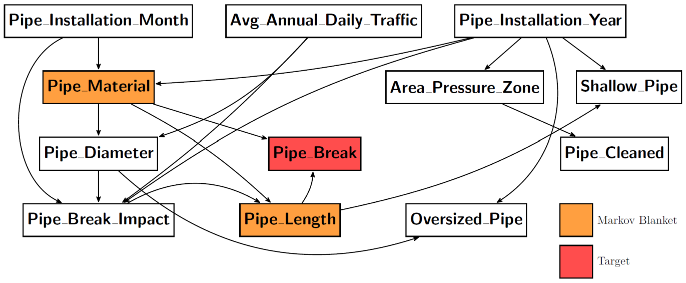

Figure 1: Subset of the Learned BN for waterpipe breakage for City of Kitchener

[Vieira *et al.*, 2021], an ensembling method as an additional model to evaluate and compare. Three XGB's survival embedding models were used to predict survival probabilities, where an xgb model was added with logistic regression(LoR), Kaplan-Meier(KM) tests on nearest neighbors, and Kaplan-Meier tests on trees. C-Index was used for evaluation.

4 Experimental Evaluation

We now describe how the aforementioned U^2RPSA approach is applied on a publicly available datasets of the water main pipe distribution. We compare the performance of our approach to state-of-the-art baselines and discuss the strengths and weaknesses of our approach. Specifically we compare the risk prediction results with Omar et al. [Omar *et al.*, 2023] which to the best of our knowledge is the state of the art of water main breakage risk prediction.

Data source, Datasets and Preprocessing: In City of Kitchener datasets [City of Kitchener, b] and [City of Kitchener, a], there are 15931 samples and 27 features in the former, and 2809 samples and 37 features in the latter. By merging the Kitchener datasets based on unique pipe identifiers, the resulting table has 64 features. After removing null values, we were left with 16,958 samples and 21 features. Categorical variables were quantified using one-hot encoding. The next step was to perform a 2-way ANOVA test, that would help us in identifying features that were statistically significant for the target. Those features whose p-value was less than 0.05 were filtered out, and the final dataset consisted of 32 features, including the target, that was labeled as '1' for break, and '0' otherwise. Lastly, the continuous-valued features were normalized by removing the mean and scaling to unit variance.

Baseline: In Omar et al.'s model, Five features out of thirty seven were selected for training based on the correlation analysis done by the authors. Feature selection using correlation coefficients was performed such that those independent variables showing correlation with the target were chosen. This narrowed the feature space to age, material, condition score, and the criticality score of the pipe, along with the average annual daily traffic the road above the pipe had to bear. The condition score here refers to a number from 0 to 10. Lower condition score means that the pipe is more prone to breakage. Criticality here refers to the impact of the damage caused to the civil infrastructure if a pipe were to break.

Evaluation: In Step 1 of U^2RPSA, we use the Bayesian Information Criterion (BIC), incorporating log-likelihood and regularization terms. The best model discerns statistical correlations in city water infrastructure data. Bayes net structures elucidate conditional probability with evidence variables, aiding interpretable inference. As observed in Figure 1, the Markov blanket consists of $Pipe\_Length$, and $Pipe\_Material$, illustrating their relationship with the $Pipe\_Break$ target outcome. The directed graph in the figure is a subset of the structure that was learned and visualized using the BN Inspector tool.

In Step 2, we applied a suite of ML algorithms to predict the risk of a pipe breakage. It was observed that the feature of $Condition\_Score$ was a 'leaky' variable. Henceforth, all the evaluation results and inferences that followed were based on models that were trained on datasets excluding the pipe's condition score feature. As this is a classification problem on whether a pipe will break or not, the comparison was based on key evaluation metrics like accuracy, precision, recall, f1-score, and area under the Receiving Operating Characteristic curve. Comparing the results of our approach with Omar et al. under the same conditions, we note a marked improvement in the classification performance in the case of some of the cases. Figures 2 and 3 describe the performance of 6 models on the F1 score and Area under the ROC curve. The red bars are the results of our approach, and the green bars are the results of the replication of the Omar et al. approach. In both charts, 4 out of 6 classifiers perform better with our approach than the replicated. As seen from the graphs, XGBoost outperforms all the other classification methods on all the key metrics. The replication of Omar et al. consisted of models that underwent randomized cross-validation for selecting the best set of hyperparameters. The 'I's on top of each bar represent the standard deviation around the mean. This can also be seen in the table 1, where the mean value was reported as the main score and the standard deviation error rate is in the parenthesis. In the AUROC chart, wherein for 3 models out of 6, our scores have a lower bound that is higher than the replicated method's upper bound. Our approach overcomes the limitation of data leakage, and shows better scores than the baseline, excluding the variable that biases the results.

The best-performing classifier was XGBoost, and was additionally used to identify the importance of the features. The top 2 features were the $Pipe\_Material$ and $Pipe\_Length$, with scores of 0.21 and 0.13 respectively. This implies that these two features provided approximately 35% of the valuable information to the decision trees within the learned model. This aspect can also be validated by observing the learned Bayesian Network of the dataset from the previous stage, where the Markov blanket of the target outcome includes $Pipe\_Material$ and $Pipe\_Length$, the 2 most important features of the XGBoost model.

For Step 3, the evaluation was done based on the Concordance index (C-index). It quantifies the proportion of cor-

	Accuracy (±SD)		AUROC (±SD)		F1 (±SD)	
Model	**Alg 1**	**Alg 2**	**Alg 1**	**Alg 2**	**Alg 1**	**Alg 2**
DT	0.91 (±0.02)	0.89 (±0.01)	0.76 (±0.06)	0.8 (±0.1)	0.59 (±0.14)	0.52 (±0.14)
NB	0.51 (±0.16)	0.85 (±0.03)	0.83 (±0.08)	0.84 (±0.09	0.34 (±0.08)	0.53 (±0.09)
KNN	0.91 (±0.01)	0.85 (±0.01)	0.86 (±0.09)	0.73 (±0.03)	0.59 (±0.15)	0.37 (±0.06)
LoR	0.91 (±0.02)	0.85 (±0.01)	0.9 (±0.02)	0.8 (±0.05)	0.56 (±0.14)	0.0 (±0.01)
RF	0.82 (±0.03)	0.9 (±0.01)	0.89 (±0.03)	0.88 (±0.05)	0.52 (±0.07)	0.56 (±0.13)
XGB	0.92 (±0.01)	0.9 (±0.01)	0.93 (±0.01)	0.9 (±0.02)	0.62 (±0.15)	0.58 (±0.14)

Table 1: Model Performance Metrics - Algo 1:- U^2RPSA, Algo 2:- Omar et al. replicated

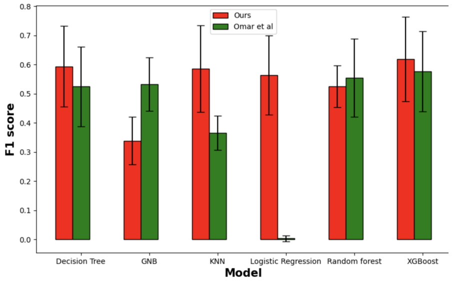

Figure 2: **Step 2 -** F1 Score

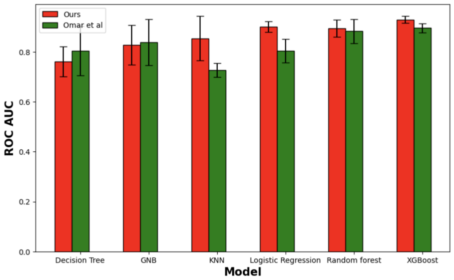

Figure 3: **Step 2 -** AUROC

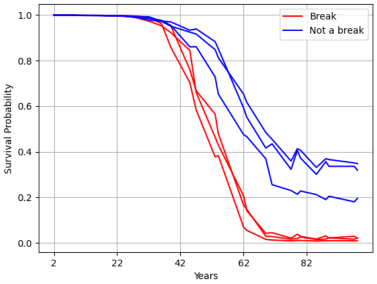

Figure 4: **Step 3 -** Survival decline

rectly ordered pairs of subjects according to their predicted risk scores such that a C-index of 0 indicates that the model's predictions are entirely incorrect, while a score of 1 indicates perfect predictive accuracy. This metric helps us understand the model's ability to rank pipes in terms of their likelihood of experiencing a break. Four models were compared: Cox-PHM, XGB+LoR model, XGB+KM on nearest neighbors model, and XGB+KM on the trees model with bootstrap meta-estimation. The C-index score were 0.7625, 0.8084, 0.7922, 0.7644 respectively. The XGB + LoR model with the best C-index score of 0.8084 has the highest predictive accuracy of survivability of the four models.

We then applied the XBG + LoR model on the test set which has 3326 rows, out of which 427 had witnessed a break. Table 2 shows some of the sample results from the test set where the column names are the years and the values of those columns are the survival probabilities for each of the six sample pipes. The six rows in the table are thus examples of the survival function prediction probabilities. The first three are for instances where the pipe breaks, and the next three are instances of those that did not break.

Figure 4 captures the survival prediction probability with respect to the number of years for a sample of pipes (six in our case) that do break (indicated by red lines) and those that do not (indicated by blue lines) in the test data. It can be observed that the survival chance of a pipe reduces after about 39 years if it was labeled as a break (red line) in the test set. Whereas, pipes that did not break showed a slow reduction in the survival probability. Also the probability of the red lines starts to decrease sharply as compared to the blue lines. The red lines represent the decreasing probability of survival, with an increase in years. Therefore, with Survival Analysis, we can estimate how long a pipe might survive given its characteristics. Overall, in regards to the pipeline, a set of pipes would have the status predicted and the survival chances estimated. This strategy aids the decision makers responsible for pipe breakage prevention to focus on those segments that need urgent attention.

Yr Br	**2**	**9**		**36**	**39**		**98**
B	0.9996	0.9989	...	0.9621	0.8580	...	0.0089
B	0.9996	0.9989	...	0.9743	0.9515	...	0.0210
B	0.9996	0.9987	...	0.9527	0.9224	...	0.0206
N	0.9996	0.9990	...	0.9735	0.9695	...	0.3184
N	0.9994	0.9986	...	0.9764	0.9541	...	0.1949
N	0.9996	0.9985	...	0.9720	0.9507	...	0.3476

Table 2: Survival prob.; Br - Break or not(B/N), Yr - Years

5 Conclusions and Future Work

The U^2RPSA approach offers a holistic solution for predicting and preventing water main breaks. As opposed to the state-of-the-art, which focuses more on augmenting the predictive abilities, we show that combining statistical techniques can take on a more global view. Future research will involve more extensive evaluation of our approach in the context of New York City water infrastructure. We will seek to incorporate real-time data and sensor technologies to boost prediction accuracy. Moreover, we are interested in studying the effect of larger scale environmental factors like climate change and urban development on our models that will facilitate water infrastructure management and examining their cascading effects across other critical infrastructure.

Acknowledgments

Our thanks to Daniel Mallia, Jangho Lee, Professors Ahearn, Sun and Stamos for their valuable input to the watermain breakage problem and the anonymous reviewers for their helpful comments. The research reported was partially supported by a PSC-CUNY 2023 grant.

References

[Akiba *et al.*, 2019] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery.

[City of Kitchener, a] City of Kitchener. Water main breaks. https://data.waterloo.ca/datasets/KitchenerGIS::water-main-breaks.

[City of Kitchener, b] City of Kitchener. Water mains. https://data.waterloo.ca/datasets/KitchenerGIS::water-mains.

[Cox, 1972] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[Demissie *et al.*, 2017] Gizachew Demissie, Solomon Tesfamariam, and Rehan Sadiq. Prediction of pipe failure by considering time-dependent factors: Dynamic bayesian belief network model. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 3(4):04017017, 2017.

[Glymour *et al.*, 2019] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10:524, June 2019.

[Kabir *et al.*, 2016] G Kabir, Solomon Tesfamariam, Jason Loeppky, and Rehan Sadiq. Predicting water main failures: A bayesian model updating approach. *Knowledge-Based Systems*, 110:144–156, 2016.

[Kimutai *et al.*, 2015] E. Kimutai, G. Betrie, R. Brander, R. Sadiq, and S. Tesfamariam. Comparison of statistical models for predicting pipe failures: Illustrative example with the city of calgary water main failure. *Journal of Pipeline Systems Engineering and Practice*, 6(4):04015005, 2015.

[Koller and Friedman, 2009] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, Cambridge Massachusetts, 2009.

[Kumar *et al.*, 2018] A. Kumar, S. A. A. Rizvi, B. Brooks, R. A. Vanderveld, K. H. Wilson, C. Kenney, ..., and R. Ghani. Using machine learning to assess the risk of and prevent water main breaks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 472–480, July 2018.

[Ley, 2023] Ana Ley. Rain Wreaks Havoc on New York's Mass Transit System. https://www.nytimes.com/2023/09/29/nyregion/nyc-flood-mta-subway.html, 2023. [Online; accessed 29-September-2023].

[Mallia, 2023] Daniel Mallia. *Towards an unsupervised Bayesian network pipeline for explainable prediction, decision making and discovery*. PhD thesis, City University of New York (CUNY), 2023.

[Nolan, 2023] Erin Nolan. Water main break in midtown manhattan floods subway system. https://www.nytimes.com/2023/08/29/nyregion/nyc-water-main-break-subway.html, 2023. [Online; accessed 29-August-2023].

[Omar *et al.*, 2023] Abdelhady Omar, Atefeh Delnaz, and Mazdak Nik-Bakht. Comparative analysis of machine learning techniques for predicting water main failures in the city of kitchener. *Journal of Infrastructure Intelligence and Resilience*, 2(3):100044, 2023.

[Scutari, 2010] Marco Scutari. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.

[Tang *et al.*, 2019] K. Tang, D. J. Parsons, and S. Jude. Comparison of automatic and guided learning for bayesian networks to analyse pipe failures in the water distribution system. *Reliability Engineering & System Safety*, 186:24–36, 2019.

[Uusitalo, 2007] Laura Uusitalo. Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, 203(3-4):312–318, May 2007.

[Vieira *et al.*, 2021] Davi Vieira, Gabriel Gimenez, Guilherme Marmerola, and Vitor Estima. Xgboost survival embeddings: improving statistical properties of xgboost survival analysis implementation, 2021.

[Á. Martínez-Codina and Garrote, 2016] D. González-Zeas Á. Martínez-Codina, M. Castillo and L. Garrote. Pressure as a predictor of occurrence of pipe breaks in water distribution networks. *Urban Water Journal*, 13(7):676–686, 2016.